Connections and symbols II

AT1
Emmanuel Dupoux
• how could a physical system ‘think’?
  – reasoning (mathematical, common sense)
  – performing complex cognitive tasks

• philosophical issue: “computational reduction”
  – reduction of unboundedly complex behavior to the combination of simple ones
    • simple set of primitive processes
    • finite set of data types
    • a finite set of operations that combine the primitive processes to make more complex ones

• Scientific issue: precise (formal, quantitative) theories of cognition
  – what computational mechanisms underly complex behaviors (like language, reasoning, etc)?
    • Symbolic IA: (sequential & deterministic) computations with symbols and rules
      – eg: Turing machines, rewrite rules, finite state automata
    • Connexionist IA: (parallel & stochastic) computation with (continuous valued) neurone-like units
      – eg: Multilevel Perceptrons, Boltzman machines

• Technological issue: usable systems (IA)
McClelland & Rumelhart’s (1986) Parallel Distributed Processing

- Cognition involves the spreading of activation, relaxation, statistical correlation.

- Represents a method for how symbolic systems might be implemented
  - Hypothesized that apparently symbolic processing is an emergent property of subsymbolic operations.

- Advantages
  - Fault tolerance & graceful degradation
  - Can be used to model learning
  - More naturally capture nonlinear relationships
  - Fuzzy information retrieval
  - bridges the gap with real neural processing
We were not prepared to produce a full-lexicon processor for different types of words. Rather, we have explored a very simple past-tense learning model consisting of two pattern associators: a phonological pattern associator and a Wickel feature pattern associator. The model is shown in Figure 1.

The basic structure of the model is as follows:

1. **Fixed Encoding Network**
2. **Pattern Associator**
   - Phonological representation
   - Wickel feature representation of root form
3. **Decoding/Binding Network**
   - Phonological representation of root form
   - Wickel feature representation of past-tense

The current guess of the corresponding past-tense form is generated as its output. The behavior of the model can be tested by giving it sentences heard in everyday experience. Rather than learning experiences to grow throughout the 200-trial period, this is the so-called U-curve that during the first 10 trials the response strength of /gAV/, /giv/, /givd/, and /givd/ grows rapidly to over .4. Unless otherwise indicated, the strengths of the correct response is stronger than the past +ed alternative. In later phases of development and the change in error type in later phases of acquisition, we have explored a very simple past-tense learning environment designed to capture the phonological structure of the root form to produce the three past-tense variants.

In Figure 5, we can see that during the first 10 trials the response strength of /gAV/ increases rapidly to over .4 while that of the regularized alternative grows rapidly to over .2. After the first 10 trials, the response strengths for the correct past tense alternatives cannot be compared with those for the base+ed alternative. Thus, we compared strengths for the correct past tense alternatives as a function of trials. The response strengths to correspond roughly to relative response probabilities for uniquely. We take these numbers as accounts for the so-called U-curve.

The percentage of correct features for regular and irregular high-frequency verbs were learned. Performance is high when only a few high-frequency regular verbs are learned. We have thus far only assumed that these response alternatives both account for a given feature, they divide the score to the binding network generation of overt responses is accomplished by a different version of the model. To examine this question, we compared strengths for the correct past tense alternative with all other past +ed alternatives. In a later section of the paper we shall discuss the pattern of past +ed errors. To turn on in the correct response again exceeds the regularized alternatives and continues to relative response probabilities.

Rumelhart & McClelland (1986)
The critique: Pinker & Mehler (1988)

- Lachter & Bever: connectionist theories are a return to associationism (Chomsky vs Skinner revisited)
- Pinker & Prince: connectionist models of the capacity to derive the past tense of English verbs is inadequate
  - rules: wug → wugged
  - exceptions: put -> put, go->went, dig->dug
- Fodor & Pylyshyn: connectionnist theories are inadequate models of language and thought
Fodor & Pylyshyn

• Position of the problem: classical theories vs connectionnism
  – Agree:
    • both classical theories & connectionism are *representationists* (they assign some ‘meaning’ to the elements – symbols or nodes)
  – Disagree
    • classical theory encode structural relationships and processes (eg, constituents, variables, rules)
    • connectionnists only encode causal relationships and processes (x causes y to fire)

• Arguments against connectionnist systems: mental representation and processes are structure sensitive
  – combinatorial semantics
    • semantics of « J. loves M. » derived from semantics of « J. », « loves » and « M. »
  – productivity
    • the list of thoughts/sentences is not finite (I can construct new thoughts with old ones)
  – systematicity
    • I construct them in a systematic way
    • eg: « x loves M. » (where x can be any proper noun)
    • eg: If I can think « J. loves M. », I can think « M. loves J. »
  – recursivity & constituent structure:
    • If I can think « P. thinks that M. is nice » I can think « J thinks that P thinks that M is nice »

-> connectionist systems have none of the above properties
Fodor & Pylyshyn (cont)

- Objections to symbolic/classical systems
  - rapidity of cognitive processes/neural speed
  - difficulty of pattern recognition/content based retrieval in conventional architectures
  - committed to rule vs exception dichotomy
  - inadequate for intuitive/nonverbal behavior
  - acutely sensitive to damage/noise (vs graceful degradation)
  - storage in classical systems is passive
  - inadequate account of gradual/frequency based application of rules
  - inadequate account of nondeterminism
  - no account of neuroscience
  → none of these arguments are valid or relevant

- CONCLUSIONS
  1. current connectionist theories are inadequate
  2. if they were to be made adequate they would be mere implementation of classical architecture
in brief

• The Fodor & Pylyshyn challenge:
  – (current) connectionist architectures fail to capture complex behaviors
  – (future) connectionist architectures are ‘mere’ implementation of symbolic architectures
• Structure of Smolensky
  – representing structures by fillers and roles
    • examples: trees, lists, etc
  – tensor products and filler/role binding (definition)
    • local, semilocal and distributed
  – unbinding (exact and self addressed)
  – capacity and graceful saturation
  – continuous and infinite structures
  – binding and unbinding networks
  – analogy between binding units and hebb weights
  – example of a stack
  – structured roles
the Smolensky response: tensor products

Paul loves Mary -> loves(Paul, Mary)
pred = loves, arg1 = Paul, arg2 = Mary
pred * loves + arg1 * Paul + arg2 * Mary

Paul loves Mary
Mary is loved by Paul

Mary loves Paul
Paul is loved by Mary

→ Problem of tensor product representations: exponential with sentence complexity

• extensions:
  – implementation of a phonological theory (Optimality Theory) in a tensor product network with energy relaxation
    • see the Harmonic Mind (Smolensky & Legendre, 2012)
  – Escaping the explosion in nb of neurons: holographic reduced representations
    • define A * B as an operation that preserves the dimensions (eg xor, circular convolution)

Plate (1995)
Elman

• structure of the paper
  – representing time
  – SRN architecture
  – xor through time
  – badiiguuu
  – word segmentation (15 words)
  – part of speech (13 categories, 29 words, 15 sentence templates)
Backprop applied

INPUT UNITS

OUTPUT UNITS

HIDDEN UNITS

CONTEXT UNITS

Figure 2. A simple recurrent network in which octivotions are copied from hidden layer to context layer on a one-for-one basis, with fixed weight of 1.0. Dotted lines represent trainable connections.

Because the patterns on the hidden units are saved as context, the hidden units must accomplish this mapping and at the same time develop representations which are useful encodings of the temporal properties of the sequential input. Thus, the internal representations that develop are sensitive to temporal context; the effect of time is implicit in these internal states. Note, however, that these representations of temporal context need not be literal. They represent a memory which is highly task- and stimulus-dependent.

Figure 7. Hierarchical cluster diagram of hidden unit activation vectors in simple sentence prediction task. Labels indicate the inputs which produced the hidden unit vectors: inputs were presented in context, and the hidden unit vectors averaged across multiple contexts.

Several points should be emphasized. First, the category structure appears to be hierarchical. Thus, "dragons" are large animals, but also members of the class of \[ -\text{human}, +\text{animate} \] nouns. The hierarchical interpretation is achieved through the way in which the spatial relations (of the representations) are organized. Representations that are near one another in the representational space form classes, while higher level categories correspond to larger and more general regions of this space. Second, it is also true that the hierarchy is "soft" and implicit. While some categories may be qualitatively distinct (i.e., very far from each other...
• extensions of Elman’s SRN
  – computational capacity of SRN

– reservoir computing
  • http://reservoir-computing.org
Conclusions

• What about the F&P Challenge?
  – tensor products are an interesting implementation/alternative to symbolic systems
  – recurrent networks could also be an alternative, but much less understood

• The hidden debate
  – innate vs learner structures (to be continued…)

Conclusions

• empirical impact of the debate
  – past tense in English
    • rule: play->played, fax->faxed
    • exceptions: sing->sang, put->put
  • Pinker & Prince (1988)
  • procedural vs declarative memory (Ullman et al, 1997; Pinker & Ullman, 2002)

Conclusions

• empirical impact of the debate (cont)
  – statistical learning vs algebraic learning in infants

  – exemplar-based versus abstract representations
    • object recognition (Biederman & Gerhardstein, 1993), face recognition, speech recognition (eg Goldinger, 1988; Johnson 1997, Pierrehumbert 2001)
Current trends

• very deep hierarchical networks
  – trained in a heavily supervised fashion
  – but, seem to correlate with human performance and monkey IT neurons


Google net, Szegedy et al, 2014

1000 categories
100,000 test
1M training
Google net: 6.67% error
Humans: 5%
Following up on recurrent networks

• LSTM
  (long short term memory)


• replaces HMM for language models


• text generation

Vinyals et al (2014)
see also Donahue et al (2014)
• playing with RNNs
  – http://karpathy.github.io/2015/05/21/rnn-effectiveness/
    (the unreasonable effectiveness of RNNs)

• memory networks
  – Weston Chopra & Bordes (2015)
  – question answering

PANDARUS:
Alas, I think he shall be come approached and the day
When little strain would be attain'd into being never fed,
And who is but a chain and subjects of his death,
I should not sleep.

Second Senator:
They are away this miseries, produced upon my soul,
Breaking and strongly should be buried, when I perish
The earth and thoughts of many states.

DUKE VINCENTIO:
Well, your wit is in the care of side and that.

Second Lord:
They would be ruled after this chamber, and
my fair nues begun out of the fact, to be conveyed,
Whose noble souls I'll have the heart of the wars.

Clown:
Come, sir, I will make did behold your worship.

VIOLA:
I'll drink it.
• attention networks
  – Xu, Ba, .. Bengio 2015

Figure 3. Examples of attending to the correct object (white indicates the attended regions, underlines indicated the corresponding word)
Neural Turing Machine – learns copy, sort

Evolutions of the symbolic systems

• probabilistic /bayesian frameworks
→ symbolic systems that learn

Le jeu des nombres

- input: nombre entre 1 et 100
- output: “oui” ou “non”

Tâche d’induction:
  - Observer quelques exemples de ‘oui’
  - Juger si des nouveaux nombres auraient pu être “oui” ou “non”.
# Le jeu des nombres

<table>
<thead>
<tr>
<th>Exemples de &quot;oui&quot;</th>
<th>Jugements de généralisation ((N = 20))</th>
</tr>
</thead>
<tbody>
<tr>
<td>60</td>
<td>Similarité diffuse</td>
</tr>
</tbody>
</table>
**Le jeu des nombres**

<table>
<thead>
<tr>
<th>Exemples de “oui”</th>
<th>Jugements de généralisation $(N = 20)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>60</td>
<td></td>
</tr>
<tr>
<td>60 80 10 30</td>
<td>Similarité diffuse</td>
</tr>
<tr>
<td></td>
<td>Règle: “multiples de 10”</td>
</tr>
</tbody>
</table>
# Le jeu des nombres

<table>
<thead>
<tr>
<th>Exemples de “oui”</th>
<th>Jugements de généralisation $(N = 20)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>60</td>
<td>Similarité diffuse</td>
</tr>
<tr>
<td>60 80 10 30</td>
<td>Règle: “multiples de 10”</td>
</tr>
<tr>
<td>60 52 57 55</td>
<td>Similarité focalisée: nombres proches de 50-60</td>
</tr>
</tbody>
</table>
Modèle Bayesien

- $H$: Espace d’hypothèse des concepts possibles
  - $h_1 = \{2, 4, 6, 8, 10, 12, \ldots, 96, 98, 100\}$ (‘‘nombres pairs’’)
  - $h_2 = \{10, 20, 30, 40, \ldots, 90, 100\}$ (‘‘multiples de 10’’)
  - $h_3 = \{2, 4, 8, 16, 32, 64\}$ (‘‘puissances de 2’’)
  - $h_4 = \{50, 51, 52, \ldots, 59, 60\}$ (‘‘nombres entre 50 et 60’’)
  - \ldots
Modèle Bayesien

- **H**: Espace d’hypothèse des concepts possibles
  - \( h_1 = \{2, 4, 6, 8, 10, 12, \ldots, 96, 98, 100\} \) ("nombres pairs")
  - \( h_2 = \{10, 20, 30, 40, \ldots, 90, 100\} \) ("multiples de 10")
  - \( h_3 = \{2, 4, 8, 16, 32, 64\} \) ("puissances de 2")
  - \( h_4 = \{50, 51, 52, \ldots, 59, 60\} \) ("nombres entre 50 et 60")
  - \ldots

- **X** = \( \{x_1, \ldots, x_n\} \): \( n \) exemples du concept C.

- Evaluer l’hypothèse h étant donné X:

\[
p(h \mid X) = \frac{p(X \mid h)p(h)}{\sum_{h' \in H} p(X \mid h')p(h')}
\]

  - \( p(h) \) [“prior”]: biais à priori
  - \( p(X \mid h) \) [“likelihood”]: information statistique dans les exemples
  - \( p(h \mid X) \) [“posterior”]: degré de confiance que h est l’extension de C.
Comparaison modèle-humain

<table>
<thead>
<tr>
<th>Exemples de “oui”</th>
<th>Jugements $(N = 20)$</th>
<th>Modèle Bayesien $(r = 0.96)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>60</td>
<td><img src="#" alt="Graph 1" /></td>
<td><img src="#" alt="Graph 2" /></td>
</tr>
<tr>
<td>60 80 10 30</td>
<td><img src="#" alt="Graph 3" /></td>
<td><img src="#" alt="Graph 4" /></td>
</tr>
<tr>
<td>60 52 57 55</td>
<td><img src="#" alt="Graph 5" /></td>
<td><img src="#" alt="Graph 6" /></td>
</tr>
</tbody>
</table>
Comparaison modèle-humain

<table>
<thead>
<tr>
<th>Exemples de &quot;oui&quot;</th>
<th>Jugements ($N = 20$)</th>
<th>Modèle Bayesien ($r = 0.93$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>16</td>
<td><img src="chart1" alt="Bar Chart" /></td>
<td><img src="chart2" alt="Bar Chart" /></td>
</tr>
<tr>
<td>16 8 2 64</td>
<td><img src="chart3" alt="Bar Chart" /></td>
<td><img src="chart4" alt="Bar Chart" /></td>
</tr>
<tr>
<td>16 23 19 20</td>
<td><img src="chart5" alt="Bar Chart" /></td>
<td><img src="chart6" alt="Bar Chart" /></td>
</tr>
</tbody>
</table>

from Josh Tennenbaum (MIT)
• Modèles linguistiques probabilistes/Natural Language processing
  • Automates finis stochastiques
  • Grammaires Context Free Probabilistes


• Learning of a syntactic/semantic parser
  • from pairs of sentence-meaning candidates


• from pairs of questions and answers (plus a database of facts)

Not covered here

• other connectionnist architectures
  – Kohonen’s maps (competitive learning) (Kohonen, 1982)
  – Adaptive Resonance Theory (Grossberg, 1976)
  – Reinforcement learning (Barto, Sutton, Anderson, 1983)

• other computational frameworks
  – more Probabilistic/Bayesian frameworks
  – Predictive Coding/Free Energy (Friston 2009)